1 Least Squares Estimation - multiple regression.

Let $\mathbf{y} = \{y_1, \cdots, y_n\}'$ be a $n \times 1$ vector of dependent variable observations. Let $\boldsymbol{\beta} = \{\beta_0, \beta_1\}'$ be the 2×1 vector of regression parameters, and $\boldsymbol{\epsilon} = \{\epsilon_1, \cdots, \epsilon_n\}'$ be the $n \times 1$ vector of additive errors. We construct the so-called *design matrix* X (dimension $n \times 2$) as follows:

$$X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ . & . \\ 1 & x_n \end{pmatrix}$$

We can now write the simple linear regression model in two ways:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \ i = 1, \cdots, n, \tag{1}$$

or equivalently

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}. \tag{2}$$

The matrix formulation easily generalizes to multiple linear regression, involving predictor variables $x_1, \cdots, x_{p-1}$. We construct the $n \times p$ design matrix X:

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & . & x_{1,p-1} \\ 1 & x_{21} & x_{22} & . & x_{2,p-1} \\ . & . & . & . & . \\ 1 & x_{n1} & x_{n2} & . & x_{n,p-1} \end{pmatrix}$$

The multiple regression can be written as

$$y_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_{p-1} x_{i,p-1} + \epsilon_i, \ i = 1, \cdots, n, \tag{3}$$

or equivalently

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{4}$$

where $\boldsymbol{\beta} = \{\beta_0, \beta_1, \cdots, \beta_{p-1}\}'$.

We use Least-Squares to fit a regression line to the data $\{\mathbf{x}_i, y_i\}_{i=1}^n$, where $\mathbf{x}_i = \{x_{i,1}, \cdots, x_{i,p-1}\}$. That is, we find the regression coefficient estimates $\widehat{\boldsymbol{\beta}}$ that minimizes the criterion

$$Q(\boldsymbol{\beta}) = (\mathbf{y} - X\boldsymbol{\beta})'(y - X\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - \mathbf{x}_i\boldsymbol{\beta})^2.$$

Taking derivatives with respect to $\boldsymbol{\beta}$, and setting these to 0, we obtain the *normal equations*:

$$\frac{dQ}{d\boldsymbol{\beta}} = -2X'(\mathbf{y} - X\boldsymbol{\beta}) = \mathbf{0} \Rightarrow$$

$$(X'X)\boldsymbol{\beta} = X'\mathbf{y} \tag{5}$$

To solve for $\boldsymbol{\beta}$ we apply the inverse of $X'X$ to both sides of equation (5) and obtain:

$$\widehat{\boldsymbol{\beta}} = (X'X)^{-1}X'\mathbf{y} \tag{6}$$

.

2 Properties

2.1 The Hat-matrix

Note, the fitted values can be written as

$$\widehat{\mathbf{y}} = X\widehat{\boldsymbol{\beta}} = X(X'X)^{-1}X'\mathbf{y},$$

where we denote the $n \times n$ matrix $X(X'X)^{-1}X'$ by H, the "Hat-matrix". The matrix H is an idempotent projection matrix:

$$HH = H \Rightarrow$$

$$X'e = X'(\mathbf{y} - \widehat{\mathbf{y}}) = X'(\mathbf{y} - H\mathbf{y}) = X'(I - H)\mathbf{y} = X'\mathbf{y} - (X'X)(X'X)^{-1}X'\mathbf{y} = 0,$$

i.e. the residuals are orthogonal to all predictor variables. In addition,

$$e'\widehat{\mathbf{y}} = ((I - H)\mathbf{y})'H\mathbf{y} = \mathbf{y}'(I - H)H\mathbf{y} = 0,$$

i.e. fitted values are orthogonal to the residuals.

2.2 Mean and Variance

$$E[\widehat{\boldsymbol{\beta}}] = E[(X'X)^{-1}X'\mathbf{y}] = (X'X)^{-1}X'X\boldsymbol{\beta} = \boldsymbol{\beta}.$$

I.e., the least-squares estimates are unbiased.

$$V[\widehat{\boldsymbol{\beta}}] = V[(X'X)^{-1}X'\mathbf{y}] = (X'X)^{-1}X'V(\mathbf{y})X(X'X)^{-1},$$

since X is not random. $V(\mathbf{y}) = \sigma^2 I$, since the errors are uncorrelated (and therefore so are the $y's$). It follows that

$$V[\widehat{\boldsymbol{\beta}}] = \sigma^2(X'X)^{-1}.$$

2.3 Interpretation

Note, $X'X \sim Cov(X)$, where the diagonal of the $p \times p$ matrix $X'X$ is the variances of the individual predictor variables (assuming x's are centered). Now, what would happen in some of the predictor variables are closely related (e.g. weight and height). If individual x's are correlated (close to linearly dependent), the $X'X$ matrix is near-singular. To solve for $\boldsymbol{\beta}$ we need to apply the inverse of $X'X$ to both sides of equation (5). If $X'X$ is near-singular this is a highly unstable operation.

What does this mean? Well, consider the regression model in equation (3). If x_1 and x_2 are closely related predictor variables, then we have no way of distinguishing between them in the regression model. Let's take the extreme example $x_1 = x_2$. If this is the case, then any combination of β_1, β_2 where $\beta_1 + \beta_2$ is constant is an equally good regression model. This extreme case is an example of an "unidentifiable" model - there is no unique best model.

The effect of this is seen in the variance of the least squares estimates. If $X'X$ is near-singular, the determinant is close to 0 and the terms in the inverse $(X'X)^{-1}$ can get very large. Therefore, the variance of the estimates $\widehat{\boldsymbol{\beta}}$ is high whenever predictor variables are correlated. For correlated predictor variables x_1, x_2 you would expect $\hat{\beta}_1$ and $\hat{\beta}_2$ to have high variance and be *negatively* correlated (since their sum $\beta_1 + \beta_2 \simeq$ constant).

In Lab 2, section 5 you can study this phenomenon via a simulation study.

2.4 Basic Inference

If we assume $\epsilon \sim N(0, \sigma^2)$, the derivation of the t-test and F-test in the multiple regression case follow from the same line of thought as the simple case.

We thus have:

- The test statistics $F_{observed} = [(SS_T - SS_E)/(p-1)]/[SS_E/(n-p)]$, where SS_T is the total sum of squares $\sum_i (y_i - \bar{y})^2$, and SS_E is the error sum of squares in the p-parameter multiple regression fit: $\sum_i (y_i - \hat{y}_i)^2$.
- Under the null, $\beta_j = 0$ for all $j = 1, \cdots, p-1$, both $SS_T/(n-1)$ and $SS_E/(n-p)$ as well as $SS_{reg}/(p-1) = (SS_T - SS_E)/(p-1)$ provide estimates for the error variance σ^2.
- Under the null, we thus expect $F_{observed}$ to be close to 1. In fact, under the null, $F_{observed}$ should come from an F-distribution with $p-1$ and $n-p$ degrees of freedom.
- We compare $F_{observed}$ to the $1-\alpha$ quantiles of the $F_{p-1,n-p}$ distribution. If $F_{observed}$ exceeds the $1-\alpha$ quantile, we reject the null at the α level, and conclude that at least 1 of $\beta_1, \cdots, \beta_{p-1}$ is different from 0.

Similarly, for inference on a single regression coefficient:

- We define the test statistic $t_{observed} = \hat{\beta}_j / SE(\hat{\beta}_j)$, where $SE(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2 diag((X'X)^{-1})_j}$ is the *standard error* of the estimate $\hat{\beta}_j$ (Remember, $V(\widehat{\boldsymbol{\beta}}) = \sigma^2(X'X)^{-1}$).
- If the true $\beta_j = 0$, $t_{observed}$ should come from a t-distribution with $n - p$ degrees of freedom.
- If the true $\beta_j \neq 0$, the test statistic will be inflated (positive or negative).
- We reject the null hypothesis if $|t_{observed}|$ exceeds the $1 - \alpha/2$ quantile of the t_{n-p}-distribution.

Caveat: if x's are correlated, then so are their estimates. In that case, testing each regression coefficient separately with a t-test can be misleading. Mathematically, you can't really tell them apart.

3 Writing a Lab Report

Some general guidelines:

1. Perform spell check! Structure the report. Use paragraphs. Don't use colloquial expressions or slang.
2. The goal of the lap report is for you to show me (convince me) that you understood *why* I asked you to do certain things. Explain what you see in the figures. Comment on the results, interpret the models. Draw conclusions, what is the take-home message?
3. Don't contradict your results, or try to guess what I want to hear. This never works. If something went wrong with the lab, or you don't understand exactly what you see - explain as best you can in an honest fashion. If you contradict your results because they're not what you expected, it only makes it look as if you didn't understand the point of the lab.
4. Finally, work independently.

3.1 Structuring a lab report

In general, I want you to follow the same structure as a research paper. However, we will skip the "Methods" section in the lab reports, though I do expect you to include this in your project reports.

I Introduction: Here you state the purpose of the lab, and if you have a main result that stands out, a main conclusion, you briefly highlight this as well (1/2-1 page).

II Methods: Skip this section in the labs. In general, a methods section should define the methods used in a stand-alone fashion (i.e. completely independently of the software implementation you used). Limitations, assumptions, and computational difficulties should be stated clearly.

III Results: This is the meat of the report. Make sure it's not a laundry list, or a diary of your time spent in the lab. If you tried lots of different things, but only a fraction of them "worked", you only need to include figures and tables relating to the methods that worked. However, it is always good to include a paragraph discussing other things you tried, and why you think they didn't work (e.g. particular data transforms, say). You can also use an Appendix section to discuss such things.

Pay careful attention to structure in this section. Separate different steps of the analysis with subsections and paragraphs (e.g. a data processing section, an initial modeling section, an inference section, etc).

Place graphs and tables in the portion of the report where they are discussed in the text, not as separate pages or in the appendix. You can use the appendix for analysis results that are not pertinent for the lab, but you still wish to discuss.

Graphs and tables should the labeled and have captions. The captions should explain fully what is in the graph (table), and include at least one sentence with a "value statement" - i.e. what story is the graph telling? Example: "Observation 12 appears to be an influential outlier."

Do not cut-and-paste software output into your report. Results from the software package should be put into tables.

Only include a graph or table if you discuss it in the body of the report, and vice versa. Be selective! Surely not all graphs you examined are equally important? Before you start writing, sit down and think about the graphs and results that lead to the main conclusion. These are the ones you should discuss in great detail and highlight in the report. Supplementary materials should not be allowed to take away focus from the main message!

IV Conclusion: Here is your chance to shine. Tell me what you think the main results are. Were there any surprises? Tell me about them.

Did the lab work raise any concerns? E.g. perhaps you had an idea about applying something we did in the lab to another type of data, and thinking about it, you're not sure that would work - can you tell me why? Perhaps you're concerned about this data set - was it too noisy, too many outliers?

Any ideas for future work - what kind of analyses would you recommend one undertakes to analyze this data set further?